Detection of Anomalous Events in a Tennis Game Using Multimodal Information

Qiang Huang* and Stephen Cox†
* University of Edinburgh, Edinburgh, UK
E-mail: qhuang23@inf.ed.ac.uk
† University of East Anglia, Norwich, UK
E-mail: s.j.cox@uea.ac.uk

***Abstract*—In the automatic analysis of a tennis game, it is important to detect some anomalous match events, such as "fault serve" and "ball out", as these events are crucial in understanding the progress of a game. Audio information can be used to detect these events, but it is unreliable, because of the acoustic mismatch between the training and the test data and interfering noise caused by spectator applause, players' yells etc. We present a framework to detect these events in which audio and visual information are used both separately and in combination. We accumulate audio evidence for anomalous events that is based on audio event classification and pitch estimation, and combine this with video evidence based on scene segmentation (itself based on audio ball-hit detection) and estimation of the ball's trajectory. To evaluate the effectiveness and robustness of our approach, we test it on three different tennis matches. Results show that our approach outperforms several audio-based baselines: the best performance is an F-score of 61% on the test data.**

I. Introduction

Sports video analysis has attracted considerable research interest during the past ten years. It is interesting both because of its rich audio-visual information content, which also has a strong inherent syntax, and because there are several useful practical applications of such analysis, such as highlight extraction[1], tactics analysis[2], computer-assisted refereeing[3]. Research in this area has also been influential in information retrieval [4], audio contents analysis [5], and tracking motion objects [6].

Our long-term goal is to study how to enable a machine to learn complex human activities by information acquisition and analysis. In our earlier work in tennis match analysis [7], we found that the use of multimodal information is essential for accurate detection of match events. Here, we focus on the detection of anomalous match events. We choose the event "ball out", which occurs whenever, during play, the ball bounces outside the permitted lines drawn on the court and brings play to a halt. In tennis, such anomalous match events are always reported by line judges, and their shouts can be heard following these events.

There have recently been several studies in the field of the content analysis using audio information. Some work [8], [9], [10], has put more focus on audio information. [8] employed a spectral clustering algorithm to discover the audio elements. [9] proposed a discriminative feature set for acoustic event detection according to approximated Bayesian accuracy. [10] built a two-stage classifier for normal and "excited" events classification. Our own previous work [12] also tried to improve the audio event detection with a hierarchical language model. However, none of this work has addressed the problem of interfering noise and the acoustic mismatch between the training and the test data, which we address here by combining audio and visual information. The paper is organised as follows: our theoretical framework is introduced in section 2, in which we describe our approaches to anomaly events detection in more detail; information about the data used in this paper is given in section 3; results and analyses are presented in section 4, and we finally summarise this paper and discuss our future work in section 5.

II. Theoretical Framework

Following our previous work [12], we define seven classes of match events ($E_i, 1 \leq i \leq 7$), which are 1. umpire's announcement, 2. commentary, 3. crowd noise, 4. line judge's shout, 5. sound of ball hit, 6. electronic beep 7. any audio event not belonging to the preceding six classes. Equation 1 shows that our aim is to identify the most likely anomalous events (E^*_{anom}) according to both audio (O^a) and visual (O^v) information.

$$E^*_{anom} = \max_{E_i} \Pr(E_i|O^a, O^v) \tag{1}$$

Audio information is exploited in two ways: the audio stream is converted to MFCCs (see Section 3 for details) and a Gaussian mixture model (GMM) is used to model each event class ($\Pr(O^a_{MFCC}|E_i)$). In addition, we use Gaussian approximations of PDFs of estimates of fundamental frequency ($F0$) extracted from detected voiced signals ($\Pr(O^a_{F0}|E_i)$). Visual information is processed to form scene models ($\Pr(O^v_{scene}|E_i)$) that are used to segment the video into two classes, "play-shot" and "non-play-shot". In visual sequences classified as "play-shot", we estimate the ball trajectory ($\Pr(O^v_{trajectory}|E_i)$). Hence equation 1 can be expanded as:

$$\begin{aligned}\Pr(E_i|O^a, O^v) &\approx \Pr(O^a|E_i) * \Pr(O^v|E_i) * \Pr(E_i) \\ &\approx \Pr(O^a_{MFCC}|E_i) * \Pr(O^a_{F0}|E_i) * \\ &\quad \Pr(O^v_{scene}|E_i) * \Pr(O^v_{trajectory}|E_i) * \Pr(E_i)\end{aligned} \tag{2}$$

where $P(E_i)$ can be viewed as a prior probability of each event class (set equal in this paper).

A. Audio likelihood based Detection

As in our previous work on the detection of the sounds of ball hits [12], we identify the line judge's shout using a standard maximum-likelihood framework by searching for the most likely audio event given the extracted MFCC sequence and the GMM event models.

To reduce the impact of acoustic mismatch, we employ a confidence measure (CM). The likelihood of each audio event class for each frame is estimated using the Gaussian mixture models of audio events built from the training-data, and the difference between highest log likelihood (LLK) and the next highest is used as a CM for that frame. The CM for an event E_i is the averaged CM of frames (f_j) within the range covered by the event.

$$CM(E_i) = \frac{1}{N}\sum_{j=1}^{N}(LLK_1(f_j^{E_i}) - LLK_2(f_j^{E_k})) \qquad (3)$$

The use of this CM provides some immunity from mismatches between the training- and test-set channel conditions: if the mismatch is high, then all the likelihoods will be low, but the overall mis-match will be cancelled out by the differencing operation, and the differences will be relatively stable within a range. A suitable threshold for the CM corresponding to a positive detection of an audio event can be determined from the training data.

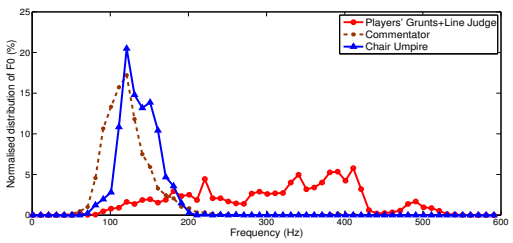

Fig. 1. Normalised distributions of F0 of the voices from commentators, chair umpire and line judges.

B. Pitch based Detection

We extract pitch information from the audio by estimating the "subharmonic-to-harmonic ratio": a detailed description of this technique can be found in [11]. Figure 1 shows the distribution of the fundamental frequency (F0) from vocalisations from the umpire, the commentators and the players and line-judges. It can be seen that the F0 of the commentators and chair umpire lies mainly within the range of 100–200 Hz, while much of the pitch extracted from the line judge calls is higher than 250 Hz. This difference enables us to coarsely locate the position of line judge calls on the sound track.

However, there is significant overlap of line judge shouts with player shouts. To effective distinguish line judges' shouts with other audio event classes and other audio interference, we build pitch based Gaussian mixture models for the line judge shout and non-line-judge audio classes, which are constructed using 3-D vectors consisting of the maximal value of F0 and its values at the start and end points of the pitch contour corresponding to the event. Audio events with larger likelihood values computed using the pitch based GMMs trained on the line judge shout are selected.

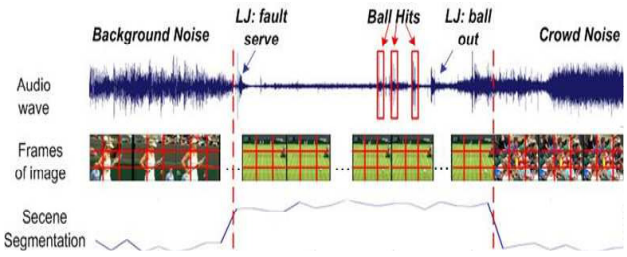

Fig. 2. Playshot scene segmentation based on the sound of ball hits and colour features

C. Scene Segmentation based Detection

Scene segmentation is based on the fact that most anomalous match events occur just before or during rallies. We divide a tennis video into two scene classes: "play-shot", covering frames in which the ball is in play, and "non-play-shot", covering other periods. Figure 2 illustrates how we identify a playshot scene using audio and visual information. The first pane shows the audio waveform, annotated with some audio events. The second pane shows (compressed) the corresponding video frames. The third pane shows the likelihood of a play-shot sequence, which peaks in the segment of the signal where the ball-hits are located.

Our approach consists of four steps:

Step1: Locate the visual frames in which ball-hit sounds are present on the corresponding section of the audio-track;

Step2: Build a visual play-shot model $\Pr(O^v|Scene_{playshot})$ using these selected visual frames ;

Step3: Compute the likelihood values of these selected frames, and average them to form a mean play-shot likelihood ($\mu - playshot$). $\mu - playshot$ is then used as a threshold for identifying play-shot frames in the video.

Step4: Compute the likelihood values of all other frames extracted from the video and discard frames whose likelihood values are less than $\mu - playshot$

The detection performance of ball hits using audio information is of the order 75%. To train a play-shot visual model, we divide each frame within a play-shot sequence into 5x5 grids. A colour histogram is computed for each of these grids and these are concatenated to generate a visual frame vector. These vectors are used to train the Gaussian mixture models of "play-shot". We finally identify which part of the video belongs to the playshot scene by selecting those frames whose computed likelihood is over a threshold determined by the averaged likelihood value of those frames selected in Step 3.

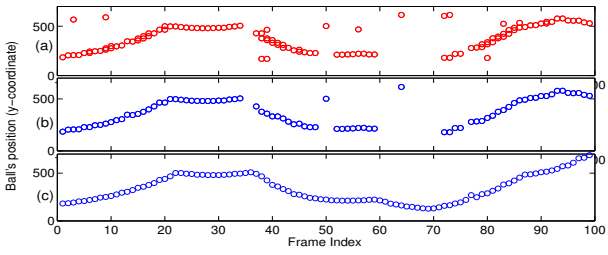

Fig. 3. Example of ball tracking

D. Ball Trajectory based Detection

To further improve the detection of the anomalous match events, we utilise ball trajectory information. We select ten visual frames ahead of the start of the detected line judges' shout because such shouts usually occur within about 0.4s of the ball bouncing, and since our visual frame-rate is 25 frames per second, $0.4 \times 25 = 10$. In practice, it is very hard to accurately locate the position where the ball bounces because of its small size, occlusion by players, and the complex visual background. Instead of attempting to locate this position accurately, as a proxy, we count (using the ball trajectory information) the number of the above frames in which the ball is located outside the lines, and divide this by 10 to obtain a probability that the ball bounced "out". Our approach hence contains two main steps:

1) determine the court region by locating all court lines
2) track the ball's motion and estimate the possibility of ball travelling outside a valid region.

To accurately find all court lines, we utilise a homography transform, described by a 3×3 matrix H, to find the mapping between points in a "virtual" tennis court template and points in the current frame. The pixel coordinate in the template is represented by a vector $[x\ y\ 1]^T$ which is multiplied by H, yielding the vector $[u\ v\ w]^T$:

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (4)$$

The final target coordinate is $(x', y') = (u/w, v/w)$. The division by w warps the coordinates properly to account for perspective foreshortening. The elements of H can be computed by mapping any four points at the corner of the court to the corresponding points in the court template. The homography transform thus enables us to further obtain the coordinates of all junctions of court lines. For a detailed description of this technique, refer to [13].

To track the ball's motion we employ the Viterbi algorithm to search for the most likely trajectory. We treat each ball candidate (b) in a frame F_t as a "state" and assume that all ball candidates in the same frame are equally likely. The observation probability of each candidate $O(b_i)$ can be obtained using the distribution (obtained from the training-data) of an actual ball's y-coordinate value in the court, namely $O(b_i) = \Pr(b_i(y))$. The transition probability between states is estimated using the distribution of the distance between two balls in two adjacent visual frames, $\Pr(D_{t,t-1}(b_i, b_j))$, again obtained from the training-data. A standard Viterbi search is used to find the most likely ball trajectory [14]. Figure 3shows an example of ball-tracking. 3(a) shows that multiple ball candidates are present in each frame, even after removing many false candidates caused by the court lines and the spectators. Also, some true candidates are missing because the ball has been blocked by players. 3(b) shows the candidates after using the Viterbi algorithm, and should be compared with 3(c), which shows the ground truth for comparison. In our experiments, the ball tracking accuracy can reach 60% (F-score).

III. Data and Experimental Set-up

We used four different tennis matches, one for training and the other three for test. Table I gives some basic information about the videos of these matches. The training data is extracted from a Wimbledon Open match, and the test matches are from the Australian Open (T1), the US Open (T2) and the WTA Paribas Open (T3). It should be noted that in these three matches, the court surfaces are all different (carpet, hard, and clay), the backgrounds are all different and the camera angles and microphone positions are all different. The soundtracks of these matches are segmented into short-time frames by a 30-ms sliding window with a 20-ms overlap. Each audio frame is be converted into a vector of 39-D MFCCs. As in our previous work[12], the seven classes of audio events are modelled with Gaussian mixture models (GMMs). Our evaluation metric is the *Fscore*:

TABLE I
Data

	Game	Type	Dur. (mins.)	# line judge
Training	Wim-08	Men-single	180	128
T1	AUS-10	Men-single	106	76
T2	US-11	Men-single	82	58
T3	WTA-12	Women-single	62	41

$$\text{Fscore} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$\text{Precision} = \frac{\#\text{correctly detected anomalous events}}{\#\text{detected anomalous events}} \quad (6)$$

$$\text{Recall} = \frac{\#\text{correctly detected anomalous events}}{\#\text{ anomalous events in the ground truth}} \quad (7)$$

A "correctly detected" anomalous event means the audio frame with a maximum likelihood value of the detected event is located within the manually annotated range of a anomalous event. Maximum likelihood values of the detected events that are not within an anomalous event range are regarded as false positives, and undetected anomalous events are false negatives.

IV. Results Analysis and Future Work

We compare seven different methods (*M1*∼*M7*) which combine the techniques introduced in section II.

M1: Audio likelihood based detection only
M2: F0 based detection only
M3: M1 + M2

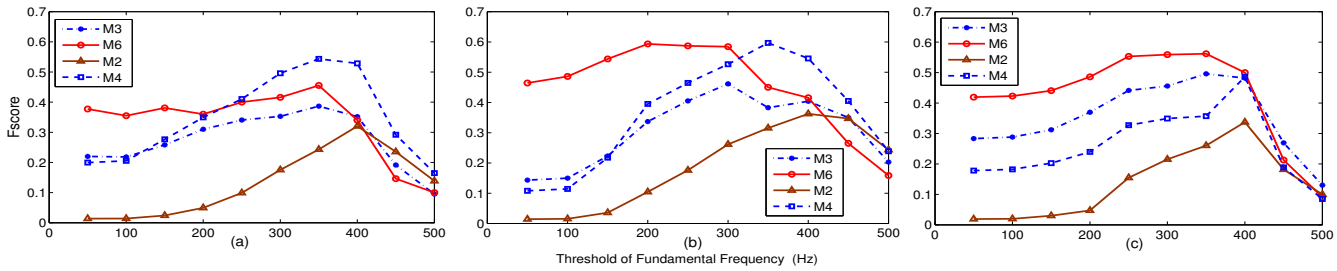

Fig. 4. Comparison of detection performances on three test matches with setting different F0 threshold: (a)-T1, (b)-T2, (c)-T3

M4: M2 + Scene Segmentation
M5: M1 + Scene Segmentation
M6: M1 + M2 + Scene Segmentation
M7: **M1 + M2 + Scene Segmentation + Ball Trajectory**

TABLE II
BEST DETECTION PERFORMANCE (FSCORE, %)

Data	M1	M2	M3	M4	M5	M6	M7
T1	21.99	32.03	38.67	54.36	37.74	45.54	**63.69**
T2	14.34	36.25	46.15	59.68	46.43	59.34	**60.16**
T3	28.32	33.73	49.59	48.60	41.94	56.18	**59.52**
Avg.	21.55	34.00	44.80	54.21	42.03	53.68	**61.12**

Table II shows the best detection performances obtained on three test matches using these seven methods. M1 (use of audio likelihoods only) is the worst-performing, due to audio mismatch between the training- and test-sets. Surprisingly, using only pitch information (M2) is superior, and using M1 and M2 in combination (M3) is considerably better than either technique on its own, because some false audio events with low F0 are removed. After employing scene segmentation, we are able to further remove false detections caused by crowd noise and commentators' voices, so that M4, M5 and M6 generally give better performance than M1, M2 and M3 (the exception being M5 compared with M3). M7 takes into account the ball's position in the court, and outperforms the other methods in all cases.

In figure 4, we compare the performances obtained using M2, M3, M4 and M6 when different fundamental frequency (F0) thresholds are used, ranging from 50 to 500 Hz. The threshold of F0 works as a pre-filter to filter out possible interference from the chair umpire speech and commentators' speech prior to applying our algorithms. We find the best performance is generally obtained within the range between 250 and 350 Hz, and the F-score is reduced considerably when the threshold is set over 400 Hz, which is what would be expected from Figure 1. M6 (audio likelihoods + F0 detection + scene segmentation) is more robust than M4 (F0 detection + scene segmentation) because the audio likelihood is a strong indicator of the position of an anomalous event.

V. SUMMARY AND DISCUSSION

We have proposed some novel techniques of integrating audio and visual information to give better detection of events in a tennis match and shown that this integration gives considerably better performance in the face of acoustic mismatch between audio soundtracks and interfering noise than using purely audio information. The techniques couple the two modalities tightly: for instance, visual scene shot segmentation is based on detection of a rally in the audio domain. They were trained and tested on matches played in different venues with different backgrounds, court-surfaces and camera and microphone positions, but perform quite robustly. Our future work will focus more on how more accurately locating ball's potion using the visual information and how more effectively fusing multimodal information.

REFERENCES

[1] Yang, Y. and Lin, S. and Zhang, Y. and Tang, S., "Highlights extraction in soccer videos based on goal-mouth detection", *Proc. 9th Int. Symposium on Signal Processing and Its Applications*, pp.1–4, 2007.
[2] Zhu, Guangyu and Huang, Qingming and Xu, Changsheng and Rui, Yong and Jiang, Shuqiang and Gao, Wen and Yao, Hongxun, "Trajectory based event tactics analysis in broadcast sports video", *Proc. 15th Int. conf. on Multimedia*, pp.58–67, 2007.
[3] Roke Manor Research Ltd, "Hawk-Eye", *http://www.bbc.co.uk/pressoffice/pressreleases/stories/2-003/06_june/10/hawk_eye.shtml*.
[4] Fleischman, M. and Roy, D., "Unsupervised Content-Based Indexing of Sports Video Retrieval", *9th ACM Workshop on Multimedia Information Retrieval (MIR)*, Augsburg, Germany, 2007.
[5] Huang, Q. and Cox, S., "Inferring the Structure of a Tennis Game using Audio Information", *IEEE Transactions on Audio, Speech and Language Processing*, vol 19(7):1925–1937, 2011.
[6] Ren, J. and Orwell, J. and Jones, A. G. and Xu, M., "Tracking the soccer ball using multiple fixed cameras", *Computer Vision and Image Understanding*, vol 113:633–642, 2009.
[7] Huang, Q., Cox, S., "Improved Detection of Ball Hit Events in a Tennis Game Using Multimodal Information", IEEE International Conference on Auditory and Visual Speech Processing, 2011.
[8] Cai, R. and Lu, L., Zhang, H.-J., and Cai, L.-H., "Highlight sound effects detection in audio stream", *in Proceedings of ICME*, pp.37–40, 2003.
[9] Zhuang, X. and Zhou, X. and Huang, T and Hasegawa-Johnson, M., "Feature analysis and selection for acoustic event detection", *in Proceedings of ICASSP*, pp.17–20, 2008.
[10] Atrey, P. and Maddage, N. and Kankanhalli, M., "Audio Based Event Detection for Multimedia Surveillance", *in Proceedings of ICASSP*, pp.813–816, 2006.
[11] Sun, X., "Pitch Determination And Voice Quality Analysis Using Subharmonic-To-Harmonic Ratio", In Proceedings of ICASSP, pp.200–203, 2002.
[12] Huang, Q. and Cox, S., "Hierarchical Language Modeling for Audio Events Detection in a Sports Game", In Proceedings of ICASSP, pp.2286–2289, Dallas, USA, 2010.
[13] Hartley, R. I. and Zisserman, A., *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521623049.
[14] Rabiner, L., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *in Proceedings of the IEEE*, pp. 257-286, 1989.